Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training

Chuhao Wu, **Jackie Cha**, Purdue University, West Lafayette, Indiana, USA, **Jay Sulek**, Indiana University, Indianapolis, USA, **Tian Zhou**, Purdue University, West Lafayette, Indiana, USA, **Chandru P. Sundaram**, Indiana University, Indianapolis, USA, and **Juan Wachs**, **Denny Yu**, Purdue University, West Lafayette, Indiana, USA

Objective: The aim of this study is to assess the relationship between eye-tracking measures and perceived workload in robotic surgical tasks.

Background: Robotic techniques provide improved dexterity, stereoscopic vision, and ergonomic control system over laparoscopic surgery, but the complexity of the interfaces and operations may pose new challenges to surgeons and compromise patient safety. Limited studies have objectively quantified workload and its impact on performance in robotic surgery. Although not yet implemented in robotic surgery, minimally intrusive and continuous eye-tracking metrics have been shown to be sensitive to changes in workload in other domains.

Methods: Eight surgical trainees participated in 15 robotic skills simulation sessions. In each session, participants performed up to 12 simulated exercises. Correlation and mixed-effects analyses were conducted to explore the relationships between eye-tracking metrics and perceived workload. Machine learning classifiers were used to determine the sensitivity of differentiating between low and high workload with eye-tracking features.

Results: Gaze entropy increased as perceived workload increased, with a correlation of .51. Pupil diameter and gaze entropy distinguished differences in workload between task difficulty levels, and both metrics increased as task level difficulty increased. The classification model using eye-tracking features achieved an accuracy of 84.7% in predicting workload levels.

Conclusion: Eye-tracking measures can detect perceived workload during robotic tasks. They can potentially be used to identify task contributors to high workload and provide measures for robotic surgery training.

Application: Workload assessment can be used for real-time monitoring of workload in robotic surgical training and provide assessments for performance and learning.

Keywords: perceived workload, eye movements, robotics and telesurgery, simulation training, statistics and data analysis

Address correspondence to Denny Yu, School of Industrial Engineering, Purdue University, 315 Grant Street, West Lafayette, IN 47906, USA; e-mail: dennyyu@purdue.edu.

HUMAN FACTORS
Vol. XX, No. X, Month XXXX, pp. 1–22
DOI: 10.1177/0018720819874544
Article reuse guidelines: sagepub.com/journals-permissions

INTRODUCTION

Compared with traditional open surgery, minimally invasive surgery (MIS) offers potential benefits of smaller incisions, reduced infection risks, decreased postoperative pain, and shortened patient recovery time (Fuchs, 2002; Verhage, Hazebroek, Boone, & Van Hillegersberg, 2009). Despite benefits, early MIS techniques like laparoscopic surgery have been observed to increase mental and physical workload (Berguer, Chen, & Smith, 2003; Berguer, Forkey, & Smith, 2001; Hemal, Srinivas, & Charles, 2001; Yu, Lowndes, Thiels, et al., 2016) due to limitations in tactile sensation, video displays, interface design, and the disconnection of separating the surgeons' hands from target organs (Ballantyne, 2002; Hamad & Curet, 2010; Lowndes & Hallbeck, 2014; Yu, Lowndes, Morrow, et al., 2016).

Advances in robotic surgical systems have the potential to address some of the ergonomic limitations observed in laparoscopic surgery (Moorthy et al., 2004; Yu et al., 2017) by providing increased dexterity, adjustable console positions, and stereoscopic visualization (Lanfranco, Castellanos, Desai, & Meyers, 2004). Yet, mental workload in robotic surgery may be a greater concern due to increased technique complexity, unique interfaces, and the disconnection with the surgical team (Catchpole et al., 2019; Weber, Catchpole, Becker, Schlenker, & Weigl, 2018; Yu et al., 2017). For example, similar to laparoscopic surgery, flow disruptions in robotic surgery have been observed to occur frequently, and disruption severity has been associated with increased self-reported workload ($\rho = .34$) (Blikkendaal et al., 2017; Weber et al., 2018). The lack of tactile feedback is another known

disadvantage that could increase surgeon workload (Talamini, Chapman, Horgan, & Melvin, 2003; Wottawa et al., 2016) and lead to adverse surgery outcomes (Hubens, Ruppert, Balliu, & Vaneerdeweg, 2004). These new challenges necessitate quantifying and monitoring workload in robotic surgery training.

Several studies have attempted to objectively measure surgeons' workload during robotic surgery. Physical workload has been measured using surface electromyography and motion tracking sensors (Lee et al., 2014; Yu et al., 2017; Zihni, Ohu, Cavallo, Cho, & Awad, 2014). Measurement of perceived workload is more limited and has primarily focused on self-reported methods, for example, National Aeronautical and Space Administration Task Load Index (NASA-TLX; Lee et al., 2014) and Surgery Task Load Index (SURG-TLX; Moore et al., 2015). These measures have been successful in distinguishing mental workload between surgical techniques, team roles, and experience level. However, subjective approaches have potential bias (e.g., inter-subject variability and the ability to self-assess), disrupt the surgical task, and are available at the completion of the case when they are typically administered (Carswell, Clarke, & Seales, 2005; Miller, 2001; Young, Brookhuis, Wickens, & Hancock, 2015). Continuous and objective measures are needed to reliably detect specific events that increase perceived workload and to provide feedback to enhance learning.

With advances in wireless sensors and signal analytics, physiological measures are becoming more feasible in the operating room and can provide objective approaches to continuously monitor surgeons' workload without interfering intraoperative work (Dias, Ngo-Howard, Boskovski, Zenati, & Yule, 2018; S. Liu et al., 2018; Yu et al., 2017). The relationship between physiological measures and mental workload has been published in many domains. Examples of physiological measures include pupillometry, blink rate, heart rate variability (HRV), and electroencephalograms (EEGs). Applications of EEG to surgery workload are still nascent, and preliminary works have shown that EEG metrics correlated with objective performance and perceived workload during robotic procedures (Guru, Esfahani, et al., 2015; Guru, Shafiei, et al., 2015). However, the extensive setup time, intrusive setup procedure, and susceptibility to motion/muscle artifacts have limited EEG's application and reliability in the fast-paced and dynamic surgical environment (Ayaz et al., 2012; Cao, Chintamani, Pandya, & Ellis, 2009; Miller, 2001). Heart rate sensors are easier to implement and have been frequently used to infer workload (Moore et al., 2015; Roscoe, 1993). However, emotional stimulus and physical workload could also increase heart rate (Jorna, 1992, 1993), and many studies have noted that HRV might not be sensitive enough for measuring mental workload (Gabaude, Baracat, Jallais, Bonniaud, & Fort, 2012; Nickel & Nachreiner, 2003).

Similar to the aforementioned physiological measures, eye-tracking metrics have also shown strong associations with perceived workload in other domains (Beatty, 1982; de Greef, Lafeber, van Oostendorp, & Lindenberg, 2009; Marquart, Cabrall, & de Winter, 2015). With advances in wireless and wearable sensors, this approach may address some of the usability and reliability concerns of the other physiological modalities. In surgery, eye tracking has seen growing applications in training and evaluation (Henneman, Marquard, Fisher, & Gawlinski, 2017; Tien et al., 2014). These studies showed that gaze patterns differentiated between expert and novice surgeons (Khan et al., 2012; Wilson et al., 2010) and recommended projecting experts' gaze patterns to trainees to improve their performance and accelerate the learning process (Chetwood et al., 2012; Wilson et al., 2011).

Preliminary works have also applied several eye-tracking metrics to measure surgical workload. For example, peak pupil size increased with task difficulty while novices transported rubber objects over dishes with different target sizes and distances (Zheng, Jiang, & Atkins, 2015). Lower blink frequency range was associated with higher NASA-TLX ratings during simulated laparoscopic tasks (Zheng et al., 2012). In addition, blink rate was higher for experts than novices during the cutting phase of simulated microsurgery although it did not vary for any of the other phases (Bednarik, Koskinen, Vrzakova, Bartczak, & Elomaa, 2018).

However, these studies were limited to basic skills tasks and laparoscopic techniques. The accuracy of eye-tracking measures for robotic tasks with more complex interfaces remains unknown. Research is needed to determine the impact of robotic interfaces and high technical complexity of telesurgery on eye-tracking technology's implementation and its ability to predict workload.

This initial study aims to explore the relationship between perceived workload and eye-tracking metrics in robotic surgical tasks. Workload is manipulated by task difficulty, as perceived workload tends to increase with increased task demand (Marinescu et al., 2018; Miyake, 2001). We hypothesize that (1) eye-tracking metrics can predict trainees' perceived workload and (2) eye-tracking metrics are sensitive to task difficulty levels.

MATERIALS AND METHODS

Participants

This study was reviewed by the university's institutional review board. The study population was surgical trainees who participated in robotic skills training (i.e., limited previous robotic experience). Eight surgical trainees from a large academic medical school were recruited voluntarily. All of the participants were right-hand dominant, four were female, and the mean (±standard deviation) age was 26 (± 1.6) years. None had prior clinical robotics experience. They performed robotics tasks (described later) periodically over the course of 4 months.

Robotic System and Tasks

The da Vinci Surgical System (dVSS; Intuitive Surgical, Inc., Sunnyvale, CA) was used at times when it was not needed for clinical procedures. The system consisted of a surgeon console with controls (e.g., foot pedals, master controls, and controls to adjust positioning) and tele-surgical robotic arms. The console also included a widely used simulation software (M-Sim®) provided by the da Vinci manufacturer, which enabled trainees to perform simulated exercises without physically activating the actual robotic arms. Both the console and the software were used in this study.

Tasks and difficulties were selected from the simulation software based on recommendations from the surgical education community. Interviews with experts in robotic surgery and medical education were used to select six tasks that can assess skills required to perform robotic surgery. These tasks required trainees to use camera control, endowrist manipulation, clutching, needle control, and needle driving to transfer or suture objects (Alzahrani et al., 2013; Perrenot et al., 2012). Depending on the specific task, up to three levels of difficulty were available in the simulation software, and all available levels were used in the study. A task at a certain level is referred to as an *exercise* in this paper. Preliminary task analysis based on the human processor model (Card, Moran, & Newell, 1986; Y. Liu, Feyen, & Tsimhoni, 2006) and Therbligs (Gilbreth & Kent, 1911) was conducted to briefly describe the task demands across task levels. The human processor model divides the task process into three discrete serial stages: perceptual, cognitive, and motor. For our tasks, these were translated into visual, cognitive, and manual demand of the task. Within each demand, actions were decomposed into basic motion elements defined by Therbligs. See Table A1 in the appendix for task descriptions and task demands. Task order was not randomized due to the curriculum-building nature of the training sessions, that is, simpler tasks were prerequisites of more advanced tasks. Based on the task order in previous studies (Finnegan, Meraney, Staff, & Shichman, 2012; Kenney, Wszolek, Gould, Libertino, & Moinzadeh, 2009), tasks were performed in the following order: Camera Targeting, Peg Board, Ring and Rail, Sponge Suturing, Dots and Needles, and Tubes. In each task, lower (easier) levels were presented before higher (more difficult) levels.

Data Collection

Performance data. The simulation software automatically assessed trainees' performance based on several criteria, for example, task time, economy of motion, drops, instrument collisions, excessive instrument force, instrument out of view, and master workspace range (Perrenot et al., 2012), which was summarized as an overall score (0%–100%) with higher scores

representing better performance. This overall score was recorded and used as the measurement of performance. Due to the design of the software, this overall score was displayed upon completion of each exercise, allowing the participant to see their performance score.

NASA-TLX. The NASA-TLX survey (Hart & Staveland, 1988) was used to assess perceived workload. The NASA-TLX contains six sub-dimensions of workload (mental demand, physical demand, temporal demand, performance, effort, and frustration) and each was rated on a visual analogue scale that ranged from 0 (*very low*) to 10 (*very high*). Scores from each dimension were summed to calculate the final NASA-TLX workload score, resulting in a final value between 0 and 60. Although a weighted NASA-TLX has also been used by other investigators, many studies have demonstrated a summed score as an acceptable implementation of NASA-TLX (Hart, 2006).

Eye-tracking metrics. A wearable eye-tracking system, Tobii Pro Glasses 2.0 (Tobii Technology AB, Danderyd, Sweden) was used to binocularly sample eye movements at 50 Hz. The eye-tracking device consisted of two major parts. A camera was located in the middle of the glass frame (outer side) to record the view of the scene while sensors were mounted in the inner side of the glass frame to capture eye movements and pupil diameter.

Pupil diameter and gaze points were continuously recorded by the system during sessions. Recordings were annotated using the Tobii Pro Lab Software (Tobii Technology AB) and extracted for further analysis. Four eye-tracking metrics were calculated from the raw data: pupil diameter (mean of left and right), gaze entropy, fixation duration, and percentage of eyelid closure (PERCLOS), defined as follows.

Pupil diameter. This metric was estimated by the eye-tracking system using images of the eyes. Previous work showed association between larger pupillary dilations and increased cognitive load (Beatty, 1982; Beatty & Kahneman, 1966; Granholm & Steinhauer, 2004; Palinko, Kun, Shyrokov, & Heeman, 2010; Pomplun & Sunkara, 2003).

Gaze entropy. It is an index that measured visual scanning randomness and was previously used as a measure of mental workload in aviation tasks (Harris, Tole, Stephens, & Ephrath, 1982; Tole, 1983). The rationale was that the exploration pattern became more random when workload increased, but divergent results had been reported in previous studies (Allsop & Gray, 2014; Di Nocera, Camilli, & Terenzi, 2007). It was adopted for the current study and calculated based on the Shannon entropy theory (Di Stasi et al., 2016; Shannon, 2001):

$$H_g(X) = -\sum p(x,y) \cdot \log_2 p(x,y),$$

where $p(x,y)$ was the probability of gaze falling in the $p(x,y)$. A gaze point was estimated as coordinates in relation to the two-dimensional field of view (1,920 × 1,080). Gaze entropy for an exercise was calculated based on all gaze points that were monitored during the exercise, across all possible x and y in the field of view.

Fixation duration. It is the total amount of time spent in fixations. Studies had suggested that fixation duration reflected information processing load (Morris, Rayner, & Pollatsek, 1990; Reimer, Mehler, Wang, & Coughlin, 2010) and increased as workload increased (de Greef et al., 2009; Recarte & Nunes, 2000). We scaled the absolute time to the percentage of time in the exercise duration:

$$\mathrm{FD}_{\%} = \frac{\text{Sum of fixation durations}}{\text{Exercise duration}} \times 100\%.$$

PERCLOS. In previous research, PERCLOS had been calculated as the percentage of time during which the pupils were covered by the eyelids by more than 80% of their area (Wierwille, Wreggit, Kirn, Ellsworth, & Fairbanks, 1994). Studies had shown that higher PERCLOS reflected increased fatigue and decreased vigilance (Marquart et al., 2015; Singh, Bhatia, & Kaur, 2011; Sommer & Golz, 2010). It had also been used as a machine learning feature to predict workload (Halverson, Estepp, Christensen, & Monnin, 2012; Tian, Zhang, Wang, Yan, & Chen, 2019). In this study, since the device did not support eyelid closure measurement, it was estimated by the

percentage of time duration (per exercise) where neither left pupil nor right pupil was detected. Since participants' head movements were constrained, this estimation was not confounded by participants looking away. It could be potentially confounded by missing data (lost pupil frames due to device malfunction), which was 1% for our device.

Study Procedure

This study was an exploratory study to determine the potential usage of eye tracking in robotic surgery, and a prospective observational study design was used. The number of sessions was not predetermined, and participants were observed every time they attended robotic training sessions over the 3-month study period. Data collection sessions were also scheduled based on robotic console availability. Participants were informed of the study at least 1 week in advance. Data collection was conducted when any participant confirmed attendance.

For each session, after arriving to the operating room, the participants reviewed a study information sheet and completed the demographic questionnaire. They were then fitted with the eye-tracking system. The system was calibrated at the beginning of each session. Baseline pupil diameter for the participants was collected following procedures recommended by previous work (Beatty & Lucero-Wagoner, 2000; Marshall, 2000; Mosaly, Mazur, & Marks, 2017). Specifically, each participant looked at the center of a white screen for 10 s (minimum diameter) and then a black screen (maximum diameter) for 10 s.

Instructions for basic operations of the console (e.g., functions of buttons, and foot pedals) were provided to all participants in their first session. Although they were allowed to familiarize themselves with the controls, no practice sessions on the study tasks were provided. During each task, the console would display pre-programmed messages on task goals and operations, and a researcher was present to address any questions or concerns throughout the session. In each session, participants were expected to perform 12 exercises. To maintain consistency with the trainees' curriculum and system usage schedule, the time constraint of each session was 45 min. Therefore, considering participants' skill and capability, advanced difficulty levels were not completed in the early phase of training. After completing each exercise, the participant completed a NASA-TLX survey. Eye-tracking data were continuously recorded throughout the entire session and post-processed in the Tobii Pro Lab software.

Statistical Analysis

Pupil diameter and gaze entropy were normalized using the feature scaling formula given following (Bo, Wang, & Jiao, 2006) to scale the data to the range of [0, 1], accounting for potential variation from individual differences in pupil diameter and pupil dilation. It also prevented a distortion in analysis caused by variable magnitude differences. Here, 0 denoted the minimum value for an individual and 1 denoted the maximum for an individual:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

Repeated measures correlation tests, r_{rm} (Bakdash & Marusich, 2017), were used to examine how strongly NASA-TLX ratings were associated with task performance. Similarly, they were used to test associations between eye-tracking metrics and NASA-TLX ratings. Instead of the more common Pearson correlation, r_{rm} coefficient was estimated using analysis of covariance (ANCOVA), where participant was treated as a factor level. This technique gave a more accurate estimation of the association between two variables when underlying individual factors can affect the relationship. The formula of r_{rm} was expressed in the form of sum of squares:

$$r_{rm} = \sqrt{\frac{SS_{Measure}}{SS_{Measure} + SS_{Error}}}.$$

Mixed-effects models were used to determine eye-tracking metric sensitivity to changes in task levels (difficulty). This approach accounted for random effects of subject and repeated measures by allowing varying intercept (Cnaan, Laird, &

Slasor, 1997; Dingemanse & Dochtermann, 2013). Each task was fitted by separate models, resulting in five models (Task Tubes had only one level of difficulty, therefore the effect of difficulty was not tested). Significance level for all statistical analyses was set at $\alpha = .05$. When appropriate, *p*-values were corrected using the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

Classification

To explore the joint capability of various eye-tracking features for detecting high workload, the Naïve Bayes algorithm (Friedman, Geiger, & Goldszmidt, 1997) was used. The algorithm was based on Bayesian theorem: $P(C_j \mid X) \propto P(C_j) \prod P(x_i \mid C_j)$, the probability of a certain class, given all evidence, was the product of prior probability of the class and all conditional probabilities of evidence. Distribution for each variable was constructed based on observed data. Naïve Bayes classifier has been applied in real-world tasks with demonstrated efficiency and accuracy in error detection and text classification (Amor, Benferhat, & Elouedi, 2004; McCallum & Nigam, 1998). The main advantages of this technique were the effectiveness for small datasets (Jyothi & Bhargavi, 2009) and applicability to different types of data (Domingos & Pazzani, 1997) such as those collected in this study.

Perceived workload levels were determined by categorizing the total NASA-TLX scores into either high or low workload. Although there is still much debate on what NASA-TLX threshold is considered "high workload," some studies observed that scores above 50 to 55 (out of 100) may lead to increased performance errors (Colle & Reid, 2005; Mazur et al., 2014; Mazur, Mosaly, Hoyle, Jones, & Marks, 2013; Yu, Lowndes, Thiels, et al., 2016). Therefore, in this study, scores above 30 (out of 60) were categorized as high workload. Limited studies have discussed the threshold of low workload. In our training environment, low workload may indicate that the tasks were too easy. We assumed that the distribution of workload scores (n = 168) resembled a normal distribution (Grier, 2015), and the number of low workload instances were sampled to be the same as those in the high end. Scores in the middle were not used for classification considering that they were ambiguous and may not necessarily represent either high or low workload. A *k*-fold cross-validation procedure was used for model training and testing (Hastie, Friedman, & Tibshirani, 2001). Based on sample size, three folds were performed. A confusion matrix was used to determine the accuracy and sensitivity of eye metrics in predicting workload. All analyses were conducted in R (R Core Team, 2018; RStudio Team, 2016).

RESULTS

A total of 15 sessions across all participants were collected over the study period. Two participants completed three sessions, three participants completed two sessions, and three participants one session. A total of 168 exercises were collected, including performance scores, NASA-TLX ratings, and eye-tracking features. Minimum exercises completed in a session was $n = 8$, and all participants completed each exercise at least once. For some sessions, participants did not complete all 12 exercises as explained in the "Materials and Methods" section. Average and standard deviation of exercise completion time was 194 ± 157 s. The standard deviation was large because difficult exercises took more time than easy exercises (Table 1).

Workload and Task Performance

Repeated measures correlation was used to test the association between perceived workload (NASA-TLX rating) and task performance score. The correlation between NASA-TLX ratings and performance scores across all exercises (n = 168) was –.55 ($p < .001$), indicating that when workload was perceived to be high, performance was poorer.

Focusing on the relationship between workload and performance for each task, the correlations were significant for Ring and Rail ($r_{rm} = .79$, $p < .001$) and Suture Sponge ($r_{rm} = .79$, $p < .001$). For Camera Targeting, the value was marginally significant ($p = .053$). Correlation values for all tasks are reported in Table 2 (a). Effect sizes were defined as large, medium, and small for r_{rm} at threshold .50, .30,

TABLE 1: Mean and Standard Deviation of Completion Time by Task and Level

	CT		PB		RR		SS			DN		T
Task Level	1	2	1	2	1	2	1	2	3	1	2	
M	73	139	84.7	99	42.5	317	294	265	312	237	225	270
SD	45.4	68.9	40.4	37.4	22.5	193	201	165	208	140	83.5	68.3

Note. CT = Camera Targeting; PB = Peg Board; RR = Ring and Rail; SS = Suture Sponge; DN = Dots and Needles; T = Tubes.

TABLE 2: Repeated Correlation Between NASA-TLX and (a) Performance and (b–e) Eye Metrics

	By Task						
Metrics	CT (n = 30)	PB (n = 30)	RR (n = 30)	SS (n = 40)	DN (n = 26)	T (n = 12)	All Task (n = 168)
(a) Performance							
r_{rm}	–.48	–.09	–.79	–.61	–.46	–.52	–.55
p	.053	.782	<.001	.001	.107	.383	<.001
(b) Pupil diameter							
r_{rm}	.52	.19	.58	.43	.55	.63	–.12
p	.032	.538	.014	.032	.039	.250	.221
(c) Gaze entropy							
r_{rm}	.62	.34	.76	.49	.45	–.42	.51
p	.009	.224	<.001	.014	.119	.522	<.001
(d) Fixation duration							
r_{rm}	–.20	–.53	–.11	–.03	.07	.36	.10
p	.522	.032	.736	.851	.815	.561	.261
(e) PERCLOS							
r_{rm}	.20	.70	.13	–.04	–.08	–.61	.04
p	.522	.002	.702	.851	.815	.263	.572

Note. CT = Camera Targeting; PB = Peg Board; RR = Ring and Rail; SS = Suture Sponge; DN = Dots and Needles; T = Tubes; PERCLOS: percentage of eyelid closure.

and .10, respectively (Bakdash & Marusich, 2017).

Workload and Eye-Tracking Metrics

Correlation values for other eye-tracking metrics with perceived workload are reported in Table 2 (b–e). Of the four eye-tracking metrics, only gaze entropy had significant correlation with NASA-TLX ratings ($r_{rm} = .51$, $p < .001$), indicating increases in gaze entropy increased perceived workload. Figure 1 illustrates the distribution of eye-tracking measures and workload, colored shapes representing different participants.

Task Difficulty on Eye-Tracking Measures

Mixed-effects models were used to test eye-tracking metric sensitivity to changes in task difficulty. With task goal and skill remaining consistent, the simulator increased difficulty levels by incorporating additional task requirements, which was expected to influence workload (see Table A1). Since changes in difficulty

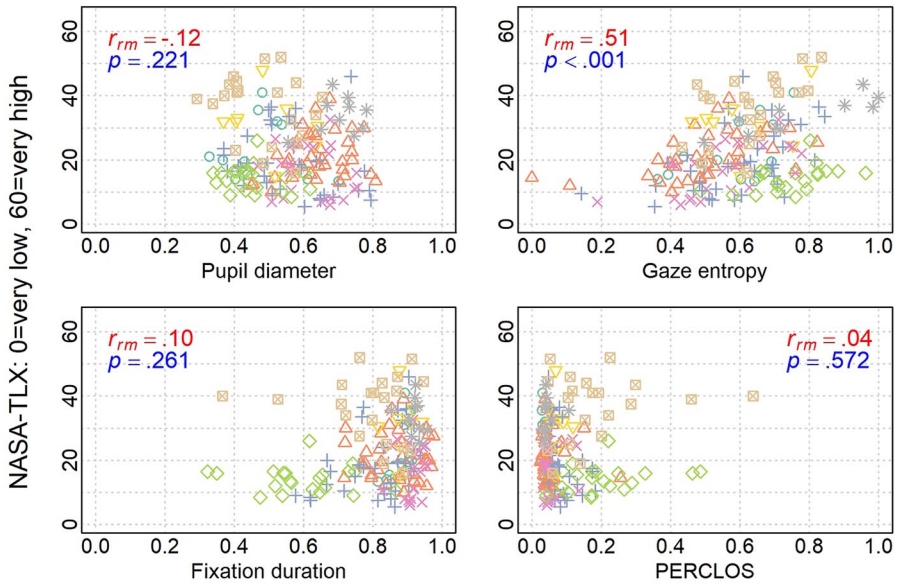

Figure 1. Distribution of eye-tracking measures over workload. Colored shapes represent different participant (only gaze entropy was significant with medium effect size). NASA-TLX = National Aeronautical and Space Administration Task Load Index; PERCLOS = percentage of eyelid closure.

TABLE 3: Mixed Models Summary for Effects of Task Level (Level 1 Is Reference Group) on Eye-Tracking Metrics

		CT	PB	RR	SS		DN
	Task Level	2	2	2	2	3	2
Pupil diameter	Coefficient	.08	.03	.12	.08	.07	.05
	p	<.001	.026	<.001	<.001	<.001	.024
	Cohen's *d*	2.38	1.04	3.02	1.49	1.40	1.20
Gaze entropy	Coefficient	.11	.05	.38	.17	.18	.03
	p	.004	.082	<.001	<.001	<.001	.427
	Cohen's *d*	1.40	0.79	3.91	1.86	2.01	0.41

Note. Level 1 was the reference level. Effect size of Cohen's *d*: small = 0.20, medium = 0.50, large = 0.80, very large = 1.20 (Sawilowsky, 2009). CT = Camera Targeting; PB = Peg Board; RR = Ring and Rail; SS = Suture Sponge; DN = Dots and Needles.

levels varied by task, each task was fitted with a model separately. Results for mixed-effects models are shown in Table 3, excluding results for fixation duration and PERCLOS, which did not reach statistical significance.

Increasing difficulty was observed to significantly increase pupil diameter for all tasks (all *p*-values <.05). The positive coefficients suggested that pupil diameters in Level 2 for all tasks were larger than that in Level 1. Level effects in tasks were very large (Cohen's *d*) except for task Peg Board. However, when there were three levels of difficulty (Suture Sponge), a Tukey post hoc test suggested that there was no difference between Levels 2 and 3 (p = .964).

TABLE 4: Mean Value of All Metrics Across Task and Level

	CT		PB		RR		SS			DN		T
Task Level	1	2	1	2	1	2	1	2	3	1	2	0
Performance	77.6	69.9	77.2	89.7	88.9	65.0	68.8	68.8	64.2	77.6	68.4	56.7
NASA-TLX	13.6	19.7	15.6	17.0	17.3	30.4	24.4	26.1	26.6	25.9	26.8	30.8
Pupil diameter	0.54	0.62	0.67	0.71	0.47	0.59	0.45	0.54	0.53	0.48	0.54	0.63
Gaze entropy	0.50	0.60	0.58	0.63	0.38	0.76	0.51	0.68	0.70	0.60	0.60	0.72
Fixation duration	0.84	0.80	0.81	0.79	0.83	0.83	0.81	0.84	0.84	0.87	0.84	0.81
PERCLOS	0.07	0.09	0.10	0.11	0.10	0.08	0.13	0.08	0.08	0.08	0.10	0.12

Note. CT = Camera Targeting; PB = Peg Board; RR = Ring and Rail; SS = Suture Sponge; DN = Dots and Needles; T = Tubes; NASA-TLX= National Aeronautical and Space Administration Task Load Index; PERCLOS = percentage of eyelid closure.

For gaze entropy, a significant effect of difficulty level was observed in the following tasks: Camera Targeting, Ring and Rail, and Suture Sponge. Based on Cohen's *d*, effects were large in all of the three tasks. The positive coefficients suggested that gaze entropy in Level 2 was greater than that of Level 1. Gaze entropy between Levels 2 and 3 in task Suture Sponge was not significantly different. Mean value of all metrics are reported in Table 4 by task and difficulty level. NASA-TLX ratings were higher in higher level of difficulty.

Workload Classification

There were 43 high workload instances with NASA-TLX scores above 30, which is the 75% quantile. The same number of instances (43) at the lowest end was labeled as low workload, which had values below or equal to 14.5 (25% quantile). Using the Naïve Bayes model, nine features were included to classify low/high workload: two demographic features (participant gender and trainee level [medical student/surgical resident]) and seven eye-tracking features (left/right pupil diameter mean, left/right pupil diameter standard deviation, gaze entropy, fixation duration, and PERCLOS). Average precision of eye-tracking measures in predicting workload was 82.8% and average classification accuracy was 84.7%. The confusion matrix for the three-fold cross-validations is presented in Table A2 in the appendix.

DISCUSSION

Eye Metrics and NASA-TLX

This study investigated the relationship between eye-tracking measures and perceived workload in robotic surgery. The first hypothesis (eye-tracking metrics can predict the level of subjective workload) was tested with both correlation analyses and machine learning classification techniques.

Gaze entropy increased significantly as NASA-TLX ratings increased. Limited studies studied the impact of gaze entropy in robotic surgery, yet Di Stasi et al. (2016, 2017) showed that gaze entropy increased with laparoscopic surgical task complexity. They explained that without knowing the optimal visual exploration strategy, surgeons might follow a suboptimal approach, which caused gaze to move constantly, especially during complex tasks.

Although many studies reported increases in pupil diameter and fixation duration with increased workload, we found no significant correlations with NASA-TLX. One possible explanation may be the robotic infrastructure. The light condition inside the enclosed console was controlled and determined by the video display of the simulation. This environment differed from previous applications of these eye-tracking metrics and may have affected pupil diameter. PERCLOS had been more commonly linked to fatigue, yet also proposed as a measure for estimating workload (Halverson

et al., 2012; Tian et al., 2019). When under prolonged states of low workload, a state of drowsiness can co-occur with a state of low attentional arousal. However, PERCLOS did not distinguish between task difficulty and perceived workload in this study. In this robotic training setting where participants were actively engaged, low arousal levels were unlikely, which can explain the low mean PERCLOS values observed.

The relationship between NASA-TLX ratings and physiological measures has been long studied, yet it remains debatable which one is a better measurement of workload. For perceived workload, NASA-TLX has been more widely used and recommended as a practical and accurate way for measuring surgeons' workload (Carswell et al., 2005; Dias et al., 2018). Recent work by Matthews, Reinerman-Jones, Barber, and Abich (2015) found that many physiological measures as well as NASA-TLX ratings were sensitive to changes in workload, but their estimates were uncorrelated. They suggested that this was caused by individual differences or the failure of assuming workload as a unitary latent construct. Other studies explained that physiological methods gave more information on how individuals responded to workload instead of what was imposed on them (Cain, 2007; Meshkati, Hancock, Rahimi, & Dawes, 1995). Our results showed that gaze entropy was significantly correlated with NASA-TLX, supporting the assumption that a latent workload construct can be estimated by both subjective and physiological methods. However, there remained unexplained variability between our gaze entropy and NASA-TLX correlation, which supports the argument that workload is multi-factorial and each method measured unique information. Therefore, the machine learning classification approach was used to combine four eye-tracking measures and investigate whether they can estimate the same level of workload as the NASA-TLX does, but in a less disruptive way.

In the Naïve Bayes model, the nine features classified between low and high workload labels with an average accuracy of 84.7%. Similar work by Halverson et al. (2012) reported an accuracy range of 16%–98% using different model specifications. In Halverson's study, there were two tasks: high workload tasks and low workload tasks, where participants needed to monitor more vehicles in the high workload task. In contrast, we did not classify the different tasks, but the different levels of perceived workload from the participants using their NASA-TLX ratings. This method is relevant to our research question of perceived workload and reflects the surgeons'/trainees' capacity of dealing with task demand. Classification of workload is clinically helpful to surgical education. The eye-tracking technique is able to provide real-time feedback on trainees' workload status, and the instances of high workload, which indicate when trainees are experiencing difficulty.

Eye Metrics and Task Difficulty

The second hypothesis tested was whether eye-tracking metrics can distinguish between varying work demands due to task difficulty level. The findings generally supported the sensitivity of eye-tracking metrics for distinguishing the differences. Mixed-effects models found significant level difficulty effects on pupil diameter and gaze entropy.

The phenomenon that pupil diameter was larger under higher level of difficulty agrees with previous studies in surgical laparoscopy (Zheng et al., 2012) and other domains (Beatty, 1982; Beatty & Kahneman, 1966; Granholm & Steinhauer, 2004; Palinko et al., 2010; Pomplun & Sunkara, 2003). Results for gaze entropy support the hypothesis that visual exploration becomes less fixed (i.e., the gaze pattern becomes more random) during more complex tasks.

Gaze Behavior in Robotic Surgery

Visual search is an indispensable step in robotic surgery. The task demands in this study were consistent with live robotic surgeries, where surgeons must rely on visual cues for completing the operation. These visual cues are delivered from the camera inside the patient that captures both current tissue states and robotic arm location. This information (e.g., current locations with respect to their desired target) is critical for planning actions necessary for completing the task goals. When searching for the target, trainees need to visually locate the target and also physically move controls to reach the

target, which constitutes a source of workload. Thus, eye-tracking measures can directly provide data for understanding trainees' task performance and learning process. For example, when trainees are unfamiliar with the environment, they may not adopt the optimal scanning strategy. As the task difficulty increases, they need more glances to compensate for the suboptimal strategy. Similarly, when trainees are novice in console operations, they tend to make mistakes and need more movements to complete tasks. Therefore, quantitative eye metrics provide feedback regarding when the trainees' visual behaviors are inefficient and when they experience high workload. Instructors can personalize training tasks to help trainees learn how to process visual cues and practice specific skills before proceeding to more complex tasks.

Although promising, future work is ongoing to address the current study's limitations. For example, due to the curriculum-progression nature, task orders in this study were not randomized, which might produce order effects. In addition, the number of sessions and exercises for each participant was not controlled in the study; having a consistent number of sessions can improve analysis accuracy and contribute to the understanding of task learning curve. Although gaze entropy was a sensitive measurement in tasks that demanded exploratory visual search, it could be less reliable in other tasks. Eye metric interaction with visual skills, cognitive skills, and manual manipulation skills have not been explored in this study but could be of potential interest in further work.

Results from this study should be viewed as initial findings from an exploratory effort. One purpose of this paper is to inspire further works on glance patterns during acquisition of new robotic surgical techniques: to prompt other researchers to explore the use of glance metrics in training and assessing surgical robotics skills. Eye-tracking metrics can identify difficult phases during training and help with the curriculum design. It may also identify trainees who are experiencing unusually high workload and are in need of extra help. In the future, more complicated techniques may be used to identify high-level tasks (Lalys, Riffaud, Bouget, & Jannin, 2012) and decompose tasks and skills (Reiley & Hager, 2009), which will augment the interpretation of workload.

APPENDIX

TABLE A1: Simulated Robotic Surgical Tasks Analysis

Camera Targeting (CT)			
Level		1	2
Objective		Focus the camera on different blue spheres spread across a broad pelvic cavity	Maintain objective 1; pick up a small cylinder under the sphere and transfer it to another sphere
Procedure		1. Search for the sphere 2. Step on the pedal to activate camera moving 3. Move both robotic arms to change the camera view[a] 4. Grip both robotic claws to activate zooming 5. Move robotic arms to zoom into the sphere **Repeat 1–5 six times**	1. Search for the sphere 2. Step on the pedal to activate camera moving 3. Move both robotic arms to change the camera view 4. Grip both robotic claws to activate zooming 5. Move robotic arms to zoom into sphere 6. Release the pedal and claws 7. Move arms to reach for the cylinder 8. Grip one claw to pick up the cylinder 9. Hold the claw 10. Search for the next sphere 11. Step on the pedal to activate camera moving 12. Move both robotic arms to change the camera view 13. Grip both robotic claws to activate zooming 14. Move robotic arms to zoom into sphere 15. Release one claw to drop the cylinder **Repeat 1–16 four times**
Demand	Manual	Minimum 24 manual movements	Minimum 52 manual movements
	Visual	Minimum 6 exploratory visual searches Minimum 12 fixations	Minimum 12 exploratory visual searches Minimum 20 fixations
	Cognitive	Recognize the signal/object Plan the movement path	Recognize the signal/object Plan the movement path
Peg Board (PB)			
Level		1	2
Objective		Grasp rings on a vertical stand with the left hand and then pass them to the right hand before placing them on a peg	Same as objective 1
Procedure		1. Find the ring that is flashing *Optional* (0–*n times*):[b]	1. Search the ring that is flashing *Optional* (1–*n times*):

(continued)

APPENDIX: (continued)

		2. Step on the pedal to activate camera moving 3. Move both robotic arms to change the camera view 4. Grip both robotic claws to activate zooming 5. Move robotic arms to zoom 6. Release the pedal and claws 7. Move arms to reach for the ring 8. Grip one claw to pick up the ring 9. Move one arm close to the other 10. Release one claw 11. Grip the other claw to transfer the ring 12. Hold the claw 13. Find the flashing peg 14. Move arms to reach for the peg 15. Release the claw to drop the ring **Repeat 1–15 six times**	2. Step on the pedal to activate camera moving 3. Move both robotic arms to change the camera view 4. Grip both robotic claws to activate zooming 5. Move robotic arms to zoom 6. Release the pedal and claws 7. Move arms to reach for the ring 8. Grip one claw to pick up the ring 9. Move one arm close to the other 10. Release one claw 11. Grip the other claw to transfer the ring 12. Hold the claw 13. Search the flashing peg *Optional* (1–*n times*): 14. Step on the pedal to activate camera moving 15. Move both robotic arms to change camera view 16. Grip both robotic claws to activate zooming 17. Move robotic arms to zoom 18. Release the pedal and claws 19. Move arms to reach for the peg 20. Release the claw to drop the ring **Repeat 1–20 six times**
Demand	Manual	Minimum 54 manual movements	Minimum 114 manual movements
	Visual	Minimum 12 exploratory visual searches Minimum 18 fixations	Minimum 12 exploratory visual searches Minimum 30 fixations
	Cognitive	Recognize the signal/object Plan the movement path	Recognize the signal/object Plan the movement path
		Ring and Rail (RR)	
Level		1	2
Objective		Move a ring along a twisted metal rod	Move three colored rings along three twisted and color-matched metal rod
Procedure		1. Move arms to reach for the ring 2. Grip one claw to pick up the ring 3. Hold the claw 4. Move arms to reach for the rod	1. Move arms to reach for ring 2. Grip one claw to pick up ring 3. Hold the claw 4. Move arms to reach for the rod

(continued)

APPENDIX (continued)

		Optional (2–*n times*): 5. Move arms to let ring go through the rod 6. Release the claw to drop the ring at the end	5. Move arms to let ring go through the rod *Optional* (2–*n times*): 6. Step on the pedal to activate camera moving 7. Move both robotic arms to change the camera view 8. Grip both robotic claws to activate zooming 9. Move robotic arms to zoom 10. Release the pedal and claws *Optional* (5–*n times*): 11. Move arms to let ring go through the rod 12. Release the claw to drop the ring at the end **Repeat 1–12 three times**
Demand	Manual	Minimum 7 manual movements	Minimum 63 manual movements
	Visual	Minimum 4 fixations	Minimum 12 fixations
	Cognitive	Recognize the signal/object Plan the movement path	Recognize the signal/object Plan the movement path Plan the order of moving three rings Estimate the force and angle needed to move the ring without being impeded by rod

Suture Sponge (SS)			
Level	1	2	3
Objective	Drive needle through random targets on a deformable sponge	Same as objective 1	Same as objective 1
Procedure	1. Move arms to reach for the needle 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arms to reach for the target 6. Move arms to drive the needle into the target 7. Find the end-target	1. Move arms to reach for the needle 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arms to reach for the target 6. Move arms to drive the needle into the target 7. Find the end-target	1. Move arms to reach for the needle 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arms to reach for the target 6. Move arms to drive the needle into the target 7. Find the end-target

(continued)

APPENDIX (continued)

6		8. Move arms to drive the needle puncture through the sponge and come out of the end-target 9. Release the claw 10. Move arms to reach the end of needle 11. Grip the claw to grip the needle 12. Move arms to pull the needle out of the sponge **Repeat 1–12 10 times**	8. Move arms to drive the needle puncture through the sponge and come out of the end-target 9. Release the claw 10. Move arms to reach the end of needle 11. Grip the claw to grip the needle 12. Move arms to pull the needle out of the sponge *Optional* (0–*n times*): 13. Step on the pedal to activate camera moving 14. Move both robotic arms to change camera view **Repeat 1–14 eight times**	8. Move arms to drive the needle puncture through the sponge and come out of the end-target 9. Release the claw 10. Move arms to reach the end of needle 11. Grip the claw to grip the needle 12. Move arms to pull the needle out of the sponge *Optional* (0–*n times*): 13. Step on the pedal to activate camera moving 14. Move both robotic arms to change the camera view **Repeat 1–14 10 times**
Demand	Manual	Minimum 120 manual movements	Minimum 96 manual movements	Minimum 120 manual movements
	Visual	Minimum 12 fixations	Minimum 12 fixations	Minimum 12 fixations
	Cognitive	Recognize the signal/object Plan the movement path Estimate the force and angle needed to drive the needle and hit the vertical end-target	Recognize the signal/object Plan the movement path Estimate the force and angle needed to drive the needle and hit the vertical and diagonal end-target	Recognize the signal/object Plan the movement path Estimate the force and angle needed to drive the needle and hit the vertical and diagonal end-target

Dots and Needles (DN)		
Level	1	2
Objective	Insert a needle through several pairs of targets that have various spatial positions	Same as objective 1
Procedure	1. Move arms to reach for the needle 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arms to reach for the target 6. Move arms to drive the needle into the target	1. Move arms to reach for the needle 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arms to reach for the target 6. Move arms to drive the needle into the target

(continued)

APPENDIX (continued)

		7. Find the end-target 8. Move arms to drive the needle puncture through the pad and come out of the end-target 9. Release the claw 10. Move arms to reach the end of needle 11. Grip the claw to grip the needle 12. Move arms to pull the needle out of the sponge **Repeat 1–12 seven times**	7. Find the end-target 8. Move arms to drive the needle puncture through the end and come out of the end-target 9. Release the claw 10. Move arms to reach the end of needle 11. Grip the claw to grip the needle 12. Move arms to pull the needle out of the sponge *Optional* (0–n *times*): 13. Step on the pedal to activate camera moving 14. Move both robotic arms to change the camera view **Repeat 1–14 six times**
Demand	Manual	Minimum 84 manual movements	Minimum 72 manual movements
	Visual	Minimum 21 fixations	Minimum 18 fixations
	Cognitive	Recognize the signal/object Plan the movement path Estimate the force and angle needed to drive the needle and hit the horizontal end-target	Recognize the signal/object Plan the movement path Estimate the force and angle needed to drive the needle and hit the horizontal and diagonal end-target
		Tubes (T)	
Objective		Drive needles through fixed targets on a cylindrical deformable structure	
Procedure		1. Search for the target 2. Move arms to reach for the cylinder 3. Grip one claw to grip the edge of the cylinder 4. Hold the claw 5. Move arms to flip the cylinder 6. Find the flashing target 7. Move the other arm to reach for the needle 8. Grip the claw to pick up the needle 9. Hold the claw 10. Move arms to drive the needle through the target 11. Release the claw 12. Move arms to reach the end of needle 13. Grip the claw to grip the needle 14. Move arms to pull the needle out of the target **Repeat 1–15 eight times**	
Demand	Manual	Minimum 120 manual movements	

(continued)

APPENDIX (continued)

Visual	Minimum 8 exploratory visual searches
	Minimum 24 fixations
Cognitive	Recognize the signal/object
	Plan the movement path
	Plan the angle of holding the cylinder
	Estimate the force and angle needed to drive the needle and hit the target
	Estimate the force and angle needed to pull out the needle without hitting the cylinder

Note. This is a high-level analysis which does not fully capture the magnitude of demands due to factors like movement/search distances and directions. A more rigorous method like Queueing Network-Model Human Processor (QN-MHP) will be needed for computational purpose.
[a]Information about distance and movement direction is not included, which differs between tasks and levels.
[b]The minimum number of optional movements is based on optimal situation, which is rarely achieved in reality.

TABLE A2: Classification Confusion Matrix

The confusion matrix shows the results of *k*-fold cross-validation of Naïve Bayes classification, with 10 indices:
True positive (TP): Proportion of instances that were classified correctly as high workload
False positive (FP): Proportion of instances that were classified incorrectly as high workload
True negative (TN): Proportion of instances that were classified correctly as low workload
False negative (FN): Proportion of instances that were classified incorrectly as low workload

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\%$$

$$\text{Negative predictive value (NPV)} = \frac{TN}{TN+FN} \times 100\%$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100\%$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100\%$$

$$\text{Accuracy} = \frac{TP+TN}{P+N} \times 100\%$$

$$\text{F1 score} = \frac{2TP}{TP+FP+FN}$$

Results were reported as the mean ± standard deviation of three validations.

	Actual Class			
	High Workload	Low Workload		
Predicted class				
High workload	44.1% ± 2.2% TP	9.4% ± 5.6% FP	82.8% ± 9.5% Precision	
Low workload	5.9% ± 2.2% FN	40.6% ± 5.5% TN	87.1% ± 5.4% NPV	
	88.3% ± 4.4% Sensitivity	81.1% ± 11.2% Specificity	84.7% ± 7.7% Accuracy	0.85 ± 0.07 F1 score

ACKNOWLEDGMENTS

The authors would like to acknowledge that this work was supported in part by Walther Oncology Physical Sciences & Engineering Research Embedding Program, Purdue Center for Cancer Research, Indiana University Simon Cancer Center; and Technology Research Grant from Intuitive Surgical, Inc. The authors would like to thank all reviewers for their feedbacks.

KEY POINTS

- Workload measurement techniques in surgery are primarily subjective, but eye tracking can be a less-intrusive, continuous, and objective workload measurement technique.
- Task performance scores, NASA-TLX ratings, and eye metrics were collected. NASA-TLX was found significantly correlated with performance. Performance accounted for between 0.84% and 62.82% of variance in NASA-TLX ratings.
- Gaze entropy was positively correlated with NASA-TLX during robotic surgical tasks. Gaze entropy accounted for between 17.51% and 38.53% of variance in NASA-TLX ratings.
- Naïve Bayes Model using the eye-tracking metrics and demographic information distinguish between self-reported workload in high and low scenarios with on average 84.7% accuracy.

ORCID IDS

Chuhao Wu https://orcid.org/0000-0002-3862-560X

Jackie Cha https://orcid.org/0000-0001-8136-2094

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article: doi:10.4231/0EVK-9P12.

REFERENCES

Allsop, J., & Gray, R. (2014). Flying under pressure: Effects of anxiety on attention and gaze behavior in aviation. *Journal of Applied Research in Memory and Cognition*, 3, 63–71.

Alzahrani, T., Haddad, R., Alkhayal, A., Delisle, J., Drudi, L., Gotlieb, W., . . . Anidjar, M. (2013). Validation of the da Vinci Surgical Skill Simulator across three surgical disciplines: A pilot study. *Canadian Urological Association Journal*, *7*, E520–E529.

Amor, N. B., Benferhat, S., & Elouedi, Z. (2004). Naive Bayes vs decision trees in intrusion detection systems. In *Proceedings of the 2004 ACM Symposium on Applied Computing* (pp. 420–424). New York, NY: Association for Computing Machinery.

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, *59*, 36–47.

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, *8*, 456.

Ballantyne, G. H. (2002). The pitfalls of laparoscopic surgery: Challenges for robotics and telerobotic surgery. *Surgical Laparoscopy, Endoscopy & Percutaneous Techniques*, *12*, 1–5.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*, 276–292.

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*, *5*, 371–372.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In *Handbook of psychophysiology* (2nd ed., pp. 142–162). New York, NY: Cambridge University Press.

Bednarik, R., Koskinen, J., Vrzakova, H., Bartczak, P., & Elomaa, A.-P. (2018). *Blink-based estimation of suturing task workload and expertise in microsurgery*. Retrieved from http://cs.uef.fi/pages/bednarik/CBMS2018_Blink_authors_version.pdf

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B: Methodological*, *57*, 289–300.

Berguer, R., Chen, J., & Smith, W. D. (2003). A comparison of the physical effort required for laparoscopic and open surgical techniques. *Archives of Surgery*, *138*, 967–970.

Berguer, R., Forkey, D. L., & Smith, W. D. (2001). The effect of laparoscopic instrument working angle on surgeons' upper extremity workload. *Surgical Endoscopy*, *15*, 1027–1029.

Blikkendaal, M. D., Driessen, S. R. C., Rodrigues, S. P., Rhemrev, J. P. T., Smeets, M. J. G. H., Dankelman, J., . . . Jansen, F. W. (2017). Surgical flow disturbances in dedicated minimally invasive surgery suites: An observational study to assess its supposed superiority over conventional suites. *Surgical Endoscopy*, *31*, 288–298.

Bo, L., Wang, L., & Jiao, L. (2006). Feature scaling for Kernel Fisher discriminant analysis using leave-one-out cross validation. *Neural Computation*, *18*, 961–978.

Cain, B. (2007). *A review of the mental workload literature*. Toronto, Ontario: Defence Research and Development Canada.

Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: Software for assessing subjective mental workload. *Behavior Research Methods*, *41*, 113–117.

Card, S., Moran, T., & Newell, A. (1986). The model human processor: An engineering model of human performance. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 2, pp. 1–35). New York, NY: Wiley.

Carswell, C. M., Clarke, D., & Seales, W. B. (2005). Assessing mental workload during laparoscopic surgery. *Surgical Innovation*, *12*, 80–90.

Catchpole, K., Bisantz, A., Hallbeck, M. S., Weigl, M., Randell, R., Kossack, M., & Anger, J. T. (2019). Human factors in robotic assisted surgery: Lessons from studies "in the Wild." *Applied Ergonomics*, *78*, 270–276. doi:10.1016/j.apergo.2018.02.011

Chetwood, A. S., Kwok, K.-W., Sun, L.-W., Mylonas, G. P., Clark, J., Darzi, A., & Yang, G.-Z. (2012). Collaborative eye tracking: A potential training tool in laparoscopic surgery. *Surgical Endoscopy*, *26*, 2003–2009.

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, *16*, 2349–2380.

Colle, H. A., & Reid, G. B. (2005). Estimating a mental workload redline in a simulated air-to-ground combat mission. *The International Journal of Aviation Psychology*, *15*, 303–319.

de Greef, T., Lafeber, H., van Oostendorp, H., & Lindenberg, J. (2009). Eye movement as indicators of mental workload to trigger adaptive automation. In D. D. Schmorrow, I. V. Estabrooke, & M. Grootjen (Eds.), *Foundations of augmented cognition: Neuroergonomics and operational neuroscience* (pp. 219–228). Heidelberg, Germany: Springer.

Dias, R. D., Ngo-Howard, M. C., Boskovski, M. T., Zenati, M. A., & Yule, S. J. (2018). Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. *British Journal of Surgery*, *105*, 491–501.

Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *Journal of Animal Ecology*, *82*, 39–54.

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, *1*, 271–285.

Di Stasi, L. L., Diaz-Piedra, C., Rieiro, H., Carrión, J. M. S., Berrido, M. M., Olivares, G., & Catena, A. (2016). Gaze entropy reflects surgical task load. *Surgical Endoscopy*, *30*, 5034–5043.

Di Stasi, L. L., Díaz-Piedra, C., Ruiz-Rabelo, J. F., Rieiro, H., Sanchez Carrion, J. M., & Catena, A. (2017). Quantifying the cognitive cost of laparo-endoscopic single-site surgeries: Gaze-based indices. *Applied Ergonomics*, *65*, 168–174.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130.

Finnegan, K. T., Meraney, A. M., Staff, I., & Shichman, S. J. (2012). da Vinci Skills Simulator construct validation study: Correlation of prior robotic experience with overall score and time score simulator performance. *Urology*, *80*, 330–336.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163.

Fuchs, K. H. (2002). Minimally invasive surgery. *Endoscopy*, *34*, 154–159.

Gabaude, C., Baracat, B., Jallais, C., Bonniaud, M., & Fort, A. (2012). Cognitive load measurement while driving. In D. de Waard, K. Brookhuis, F. Dehais, C. Weikert, & S. Röttger, et al. (Eds.), *Human factors: A view from an integrative perspective* (pp. 67–80). Toulouse, France: Human Factors and Ergonomics Society.

Gilbreth, F. B., & Kent, R. T. (1911). *Motion study*. London, England: Constable.

Granholm, E., & Steinhauer, S. R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology*, *52*, 1–6.

Grier, R. A. (2015). How high is high? A meta-analysis of NASA-TLX global workload scores. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*, 1727–1731.

Guru, K. A., Esfahani, E. T., Raza, S. J., Bhat, R., Wang, K., Hammond, Y., . . . Chowriappa, A. J. (2015). Cognitive skills assessment during robot-assisted surgery: Separating the wheat from the chaff. *BJU International*, *115*, 166–174.

Guru, K. A., Shafiei, S. B., Khan, A., Hussein, A. A., Sharif, M., & Esfahani, E. T. (2015). Understanding cognitive performance during robot-assisted surgery. *Urology*, *86*, 751–757.

Halverson, T., Estepp, J., Christensen, J., & Monnin, J. (2012). Classifying workload with eye movements in a complex task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*, 168–172.

Hamad, G. G., & Curet, M. (2010). Minimally invasive surgery. *The American Journal of Surgery*, *199*, 263–265.

Harris, R. L., Tole, J. R., Stephens, A. T., & Ephrath, A. R. (1982). Visual scanning behavior and pilot workload. *Aviation, Space, and Environmental Medicine*, *53*, 1067–1072.

Hart, S. G. (2006). NASA-task load index (NASA-TLX): 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 904–908.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). Model assessment and selection. In T. Hastie, J. Friedman, & R. Tibshirani (Eds.), *Springer Series in Statistics. The elements of statistical learning* (pp. 193–224). New York, NY: Springer.

Hemal, A. K., Srinivas, M., & Charles, A. R. (2001). Ergonomic problems associated with laparoscopy. *Journal of Endourology*, *15*, 499–503.

Henneman, E. A., Marquard, J. L., Fisher, D. L., & Gawlinski, A. (2017). Eye tracking: A novel approach for evaluating and improving the safety of healthcare processes in the simulated setting. *Simulation in Healthcare*, *12*, 51–56.

Hubens, G., Ruppert, M., Balliu, L., & Vaneerdeweg, W. (2004). What have we learnt after two years working with the da Vinci robot system in digestive surgery ? *Acta Chirurgica Belgica*, *104*, 609–614.

Jorna, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology*, *34*, 237–257.

Jorna, P. G. A. M. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics*, *36*, 1043–1054.

Jyothi, S., & Bhargavi, P. (2009). Applying Naive Bayes data mining technique for classification of agricultural land soils. *International Journal of Computer Science and Network Security*, *9*, 117–122.

Kenney, P. A., Wszolek, M. F., Gould, J. J., Libertino, J. A., & Moinzadeh, A. (2009). Face, content, and construct validity of dV-Trainer, a novel virtual reality simulator for robotic surgery. *Urology*, *73*, 1288–1292.

Khan, R. S. A., Tien, G., Atkins, M. S., Zheng, B., Panton, O. N. M., & Meneghetti, A. T. (2012). Analysis of eye gaze: Do novice surgeons look at the same location as expert surgeons during a laparoscopic operation? *Surgical Endoscopy*, *26*, 3536–3540.

Lalys, F., Riffaud, L., Bouget, D., & Jannin, P. (2012). A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. *IEEE Transactions on Biomedical Engineering*, *59*, 966–976.

Lanfranco, A. R., Castellanos, A. E., Desai, J. P., & Meyers, W. C. (2004). Robotic surgery. *Annals of Surgery*, *239*, 14–21.

Lee, G. I., Lee, M. R., Clanton, T., Sutton, E., Park, A. E., & Marohn, M. R. (2014). Comparative assessment of physical and cognitive ergonomics associated with robotic and traditional laparoscopic surgeries. *Surgical Endoscopy*, *28*, 456–465.

Liu, S., Hemming, D., Luo, R. B., Reynolds, J., Delong, J. C., Sandler, B. J., . . . Horgan, S. (2018). Solving the surgeon ergonomic crisis with surgical exosuit. *Surgical Endoscopy*, *32*, 236–244.

Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queueing Network-Model Human Processor (QN-MHP): A computational architecture for multitask performance in human-machine systems. *ACM Transactions on Computer-Human Interaction*, *13*, 37–70.

Lowndes, B. R., & Hallbeck, M. S. (2014). Overview of human factors and ergonomics in the OR, with an emphasis on minimally invasive surgeries. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *24*, 308–317.

Marinescu, A. C., Sharples, S., Ritchie, A. C., Sánchez López, T., McDowell, M., & Morvan, H. P. (2018). Physiological parameter response to variation of mental workload. *Human Factors*, *60*, 31–56.

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, *3*, 2854–2861.

Marshall, S. P. (2000). *Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity*. Retrieved from https://www.lens.org/lens/patent/102-917-720-834-189

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human Factors*, *57*, 125–143.

Mazur, L. M., Mosaly, P. R., Hoyle, L. M., Jones, E. L., Chera, B. S., & Marks, L. B. (2014). Relating physician's workload with errors during radiation therapy planning. *Practical Radiation Oncology*, *4*, 71–75.

Mazur, L. M., Mosaly, P. R., Hoyle, L. M., Jones, E. L., & Marks, L. B. (2013). Subjective and objective quantification of physician's workload and performance during radiation therapy planning tasks. *Practical Radiation Oncology*, *3*, e171–e177.

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41–48). doi:10.1.1.65.9324

Meshkati, N., Hancock, P. A., Rahimi, M., & Dawes, S. M. (1995). Techniques in mental workload assessment. In J. R. Wilson & E. N. Corlett (Eds.), *Evaluation of human work: A practical ergonomics methodology* (pp. 749–782). Philadelphia, PA: Taylor & Francis.

Miller, S. (2001). *Workload measures*. Iowa City, IA: National Advanced Driving Simulator.

Miyake, S. (2001). Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology*, *40*, 233–238.

Moore, L. J., Wilson, M. R., McGrath, J. S., Waine, E., Masters, R. S. W., & Vine, S. J. (2015). Surgeons' display reduced mental effort and workload while performing robotically assisted surgical tasks, when compared to conventional laparoscopy. *Surgical Endoscopy*, *29*, 2553–2560.

Moorthy, K., Munz, Y., Dosis, A., Hernandez, J., Martin, S., Bello, F., . . . Darzi, A. (2004). Dexterity enhancement with robotic surgery. *Surgical Endoscopy and Other Interventional Techniques*, *18*, 790–795.

Morris, R. K., Rayner, K., & Pollatsek, A. (1990). Eye movement guidance in reading: The role of parafoveal letter and space information. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 268–281.

Mosaly, P. R., Mazur, L. M., & Marks, L. B. (2017). Quantification of baseline pupillary response and task-evoked pupillary response during constant and incremental task load. *Ergonomics*, *60*, 1369–1375.

Nickel, P., & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors*, *45*, 575–590.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 141–144). New York, NY: Association for Computing Machinery.

Perrenot, C., Perez, M., Tran, N., Jehl, J.-P., Felblinger, J., Bresler, L., & Hubert, J. (2012). The virtual reality simulator dV-Trainer® is a valid assessment tool for robotic surgical skills. *Surgical Endoscopy*, *26*, 2587–2593.

Pomplun, M., & Sunkara, S. (2003). *Pupil dilation as an indicator of cognitive workload in human-computer interaction*. Proceedings of the 2009 HCI International, Crete, Greece.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied*, *6*, 31–43.

Reiley, C. E., & Hager, G. D. (2009). *Decomposition of robotic surgical tasks: An analysis of subtasks and their correlation to skill*. London, United Kingdom: M2CAI Workshop. MICCAI.

Reimer, B., Mehler, B., Wang, Y., & Coughlin, J. F. (2010). The impact of systematic variation of cognitive demand on drivers' visual attention across multiple age groups. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*, 2052–2055.

Roscoe, A. H. (1993). Heart rate as a psychophysiological measure for in-flight workload assessment. *Ergonomics*, *36*, 1055–1062.

RStudio Team. (2016). *RStudio: Integrated development environment for R*. Boston, MA: RStudio, Inc.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), Article 26.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*, 3–55.

Singh, H., Bhatia, J. S., & Kaur, J. (2011). Eye tracking based driver fatigue monitoring and warning system. In *India International Conference On Power Electronics (IICPE2010)* (pp. 1–6). New York, NY: IEEE.

Sommer, D., & Golz, M. (2010). Evaluation of PERCLOS based current fatigue monitoring technologies. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 4456–4459). New York, NY: IEEE.

Talamini, M. A., Chapman, S., Horgan, S., & Melvin, W. S. (2003). A prospective analysis of 211 robotic-assisted surgical procedures. *Surgical Endoscopy and Other Interventional Techniques*, *17*, 1521–1524.

Tian, Y., Zhang, S., Wang, C., Yan, Q., & Chen, S. (2019). Eye tracking for assessment of mental workload and evaluation of RVD interface. In S. Long & B. S. Dhillon (Eds.), *Man-machine-environment system engineering* (pp. 11-17). Singapore: Springer.

Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G.-Z., & Darzi, A. (2014). Eye tracking for skills assessment and training: A systematic review. *Journal of Surgical Research*, *191*, 169–178.

Tole, J. R. S. (1983). *Visual scanning behavior and pilot workload*. Retrieved from https://ntrs.nasa.gov/search.jsp?R=19830025266

Verhage, R. J., Hazebroek, E. J., Boone, J., & Van Hillegersberg, R. (2009). Minimally invasive surgery compared to open procedures in esophagectomy for cancer: A systematic review of the literature. *Minerva Chirurgica*, *64*, 135–146.

Weber, J., Catchpole, K., Becker, A. J., Schlenker, B., & Weigl, M. (2018). Effects of flow disruptions on mental workload and surgical performance in robotic-assisted surgery. *World Journal of Surgery*, *42*, 3599–3607.

Wierwille, W. W., Wreggit, S. S., Kirn, C. L., Ellsworth, L. A., & Fairbanks, R. J. (1994). *Research on vehicle-based driver status/performance monitoring: Development, validation, and refinement of algorithms for detection of driver drowsiness. Final report*. Retrieved from https://trid.trb.org/view/448128

Wilson, M. R., McGrath, J., Vine, S., Brewer, J., Defriend, D., & Masters, R. (2010). Psychomotor control in a virtual laparoscopic surgery training environment: Gaze control parameters differentiate novices from experts. *Surgical Endoscopy*, *24*, 2458–2464.

Wilson, M. R., Vine, S. J., Bright, E., Masters, R. S. W., Defriend, D., & McGrath, J. S. (2011). Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: A randomized, controlled study. *Surgical Endoscopy*, *25*, 3731–3739.

Wottawa, C. R., Genovese, B., Nowroozi, B. N., Hart, S. D., Bisley, J. W., Grundfest, W. S., & Dutson, E. P. (2016). Evaluating tactile feedback in robotic surgery for potential clinical application using an animal model. *Surgical Endoscopy*, *30*, 3198–3209.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*, 1–17.

Yu, D., Dural, C., Morrow, M. M. B., Yang, L., Collins, J. W., Hallbeck, S., . . . Forsman, M. (2017). Intraoperative workload in robotic surgery assessed by wearable motion tracking sensors and questionnaires. *Surgical Endoscopy*, *31*, 877–886.

Yu, D., Lowndes, B., Morrow, M., Kaufman, K., Bingener, J., & Hallbeck, S. (2016). Impact of novel shift handle laparoscopic tool on wrist ergonomics and task performance. *Surgical Endoscopy*, *30*, 3480–3490.

Yu, D., Lowndes, B., Thiels, C., Bingener, J., Abdelrahman, A., Lyons, R., & Hallbeck, S. (2016). Quantifying intraoperative workloads across the surgical team roles: Room for better balance? *World Journal of Surgery*, *40*, 1565–1574.

Zheng, B., Jiang, X., & Atkins, M. S. (2015). Detection of changes in surgical difficulty: Evidence from pupil responses. *Surgical Innovation*, *22*, 629–635.

Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O. N. M., & Atkins, M. S. (2012). Workload assessment of surgeons: Correlation between NASA TLX and blinks. *Surgical Endoscopy*, *26*, 2746–2750.

Zihni, A. M., Ohu, I., Cavallo, J. A., Cho, S., & Awad, M. M. (2014). Ergonomic analysis of robot-assisted and traditional laparoscopic procedures. *Surgical Endoscopy*, *28*, 3379–3384.

Chuhao Wu is a Master's student in Industrial Engineering (IE), Purdue University. He earned a Bachelor's degree in Engineering Management from Beijing Jiaotong University in 2017.

Jackie Cha is a PhD student in IE, Purdue University. She earned a MSE in Biomedical Engineering from the University of Michigan in 2016.

Jay Sulek is a Urologist in Urology Associates, P.C. He earned a MD from University of Maryland in 2012.

Tian Zhou is a Senior Data Scientist at Boston Consulting Group. He earned a PhD in IE from Purdue University in 2018.

Chandru P. Sundaram is a Professor of Urology at Indiana University. He earned a MD from Madras Medical College in 1985.

Juan Wachs is an Associate Professor of IE at Purdue University. He earned a PhD in Industrial Engineering and Management from the Ben-Gurion University in 2008.

Denny Yu is an Assistant Professor of IE at Purdue University. He earned a PhD in Industrial and Operations Engineering from University of Michigan in 2014.

Date received: August 23, 2018
Date accepted: August 5, 2019